Lexicography and Linguistic Diversity, 6-10 September, 2016, Tbilisi, Georgia. - Tbilisi: Ivane Javakhishvili Tbilisi State University, 2016. Pp. 457-462.

The *wdlpOst* Toolset for Creating Historical Loanword Dictionaries

(Software Demonstration)

Peter Meyer
Institut für Deutsche Sprache
e-mail: meyer@ids-mannheim.de

Abstract

The *wdlpOst* dictionary writing system to be presented in this paper has been developed for the specific purposes of a lexicographical project on German loanwords in the East Slavic languages Russian, Belarusian, and Ukrainian. The project's main objectives are (i) to document those loanwords for which a cognate lexical borrowing from German is known in Polish and (ii) to establish possible borrowing pathways for these lexical items. In the first phase of the project, the collaborative client/server architecture of the *wdlpOst* system has been used for excerpting detailed lexicographical information from a large range of historical and contemporary East Slavic dictionaries, taking the entries in a large dictionary of German loanwords in Polish as a common frame of reference. For the project's second phase, the *wdlpOst* system provides innovative tooling for compiling entries of the East Slavic loanwords. Most importantly, the numerous word sense definitions for a set of cognate loanwords, as excerpted from different lexicographical sources, are mapped onto a system of newly defined cross-language word senses; in a similar vein, the phonemic and graphemic variation in the loanwords and their derivatives is captured through a tool that abstracts from dictionary-specific idiosyncrasies.
Keywords: loanword lexicography; dictionary writing system; Slavic languages

1 Overview

This paper intends to give a brief overview of the functionality of the *wdlpOst* toolset that has been developed since mid-2013 for the specific purposes of a joint lexicographical project of the Institute for the German Language, Mannheim, and the Institute of Slavic Studies at the University of Oldenburg. The project focuses on those German loanwords in the East Slavic languages Russian, Belarusian, and Ukrainian for which a cognate German loanword in Polish is known to exist. Besides lexicographical documentation, the chief endeavor of the project is to establish the possibly complex borrowing pathways for the individual German etyma, e.g. via Polish and Old Ruthenian into Ukrainian/Belarusian and from there into Russian; or via Russian into Polish; or simply as independent borrowings.

The project's lexicographical point of departure is given by the approx. 2400 entries in the *Dictionary of German Loanwords in Standard and Written Polish* (henceforth, DGLP; de Vincenz & Hentschel 2010). In the first and now almost completed *excerption* phase of the project, 19 historical and contemporary East Slavic monolingual dictionaries, including paper slips from the unpublished parts of four large multi-volume historical dictionaries, have been scanned systematically for possible

cognates of the Polish loanwords. The candidate entries found have then been excerpted in a very detailed way with the *wdlpOst* system, including phonological variants, word senses and derivative forms, all with relevant dated quotations. For the subsequent *compilation* phase of the project the *wdlpOst* system provides sophisticated tools for creating dictionary entries of the East Slavic loanwords from the wealth of excerpted lexicographical material, with a focus on establishing the various different borrowing pathways of the candidate loanwords found (cf. Meyer 2014a). A crucial task for this second phase is that of abstracting from the idiosyncrasies and the structural and historical heterogeneity of the many sources drawn upon. This includes, above all, the necessity to define, for each German etymon, the spectrum of word senses lexicographically attested for the pertaining Polish and the East Slavic loanwords, including dates of first and possibly last quotations in the individual languages.

2 Main functions of the software

2.1 Tools for excerption

The fully bilingual (German and Russian) user interface for creating and editing dictionary excerpts (henceforth, 'editor') is geared to less computer-savvy users, abstracting entirely from the underlying XML document editing process. The entire DGLP data is integrated into the editor and serves as a common frame of reference for a large number of the editor's more than 100 input sections. Amongst other things, each excerpt must be assigned to a DGLP entry with a presumably cognate Polish loanword from German; either the lemma or one of the derivatives listed in the DGLP entry must be specified as 'corresponding' to the lemma of the East Slavic entry. For each excerpted word sense definition *D*, the lexicographer must specify its *DGLP sense profile* by indicating which of the DGLP entry word senses, if any, are subsumed under *D* completely or at least partially. In later stages of the lexicographical process, the sense profiles help in establishing the word sense spectra mentioned in section 1.

The editor offers the typical functionality of modern dictionary writing systems. Some special features related to the specific needs of the project are the following: (i) drop-down menus and keyboard shortcuts assist in entering special Unicode characters of the various scripts found in East Slavic historical dictionaries; (ii) several 'restricted input modes' allow assistant workers to excerpt and input specific types of information without the possibility of messing up other excerpt parts; (iii) there are HTML overviews of the DGLP entry and the current state of the present excerpt; (iv) the system has very fine-grained live validation procedures for all resources; (v) a server-based source management system provides consistency for source sigla and dates of quotation sources; (vi) the input dialog for quotations offers sophisticated options to specify 'fuzzy' dates where exact figures are not available (such as 'last third of the 15th century'); (vii) the editor has an offline mode used at several East European dictionary editorial offices without reliable Internet connection (in this mode, excerpts are stored on the local hard disk, but can optionally be backed up to the server or sent via e-mail). Figure 1 shows a screenshot of the editor's main dialog window and the quotation editing tool.

The editor offers a server-based structured full-text search on the entire set of XML excerpt documents. For more specific research purposes a snapshot of this set can be downloaded and pre-processed locally with a single mouse click; the application then offers integrated XPath and SQL query interfaces, the latter representing the entire XML dataset as a relational database with about 40 tables.

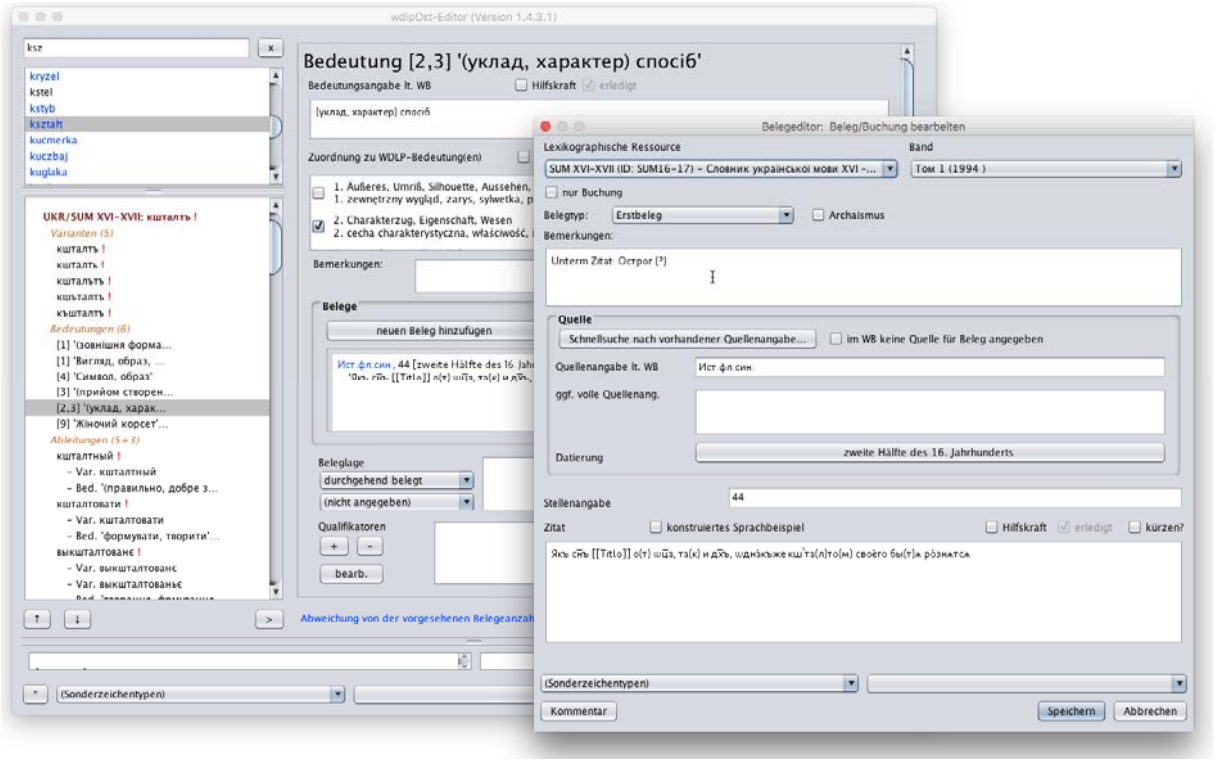

Figure 1: The main window and the quotation input dialog of the excerpt editor tool.

2.2 Tools for entry compilation

The *wdlpOst* system includes a number of innovative tools for the purposes of compiling the lexicographical output of the project. The headword of a 'target entry' is a German etymon; the entry is supposed to give a comprehensive picture of all East Slavic loanwords relating to the etymon and the possibly divergent borrowing pathways of the attested word senses and/or variants. Various compilation tools assist in creating the different sections of the target entries:[1]

- The *German etymon* with a relevant subset of its diasystematic forms and word senses. A first draft of this section can be created automatically by using data from the corresponding DGLP entry or entries.
- The *word senses* attested for the East Slavic loanwords. The overall set of word senses for a target entry is defined cross-linguistically with the 'metasense' tool described in detail in (Meyer 2015): An intuitive drag-and-drop graphical user interface (cf. Figure 2 below) is used for mapping, for each DGLP entry, the manifold excerpted sense definitions of related East Slavic words onto the manually defined word senses ('metasenses') of the target entry. A first draft of such a mapping can be constructed automatically from the DGLP sense profile annotations mentioned in 2.1. For each word sense, the first and the last or most recent quotation must be selected among the excerpted quotations. The system proposes candidate quotations on the basis of the excerpted dates.
- A table of *semantic fields*, slightly adapted from the set used in Haspelmath & Tadmor 2009, that shows, for all five languages involved (German, Polish, Russian, Ukrainian, Belarusian), which of the fields are relevant to at least one of the 'metasenses' ascribed to a word listed for this language in the entry. A draft version of the grid can be computed automatically from the

[1] Note that the following description of the target entry sections is preliminary and incomplete.

DGLP sense profiles and an already existing annotation of the DGLP dataset.

- The *(mostly phonemic) variants* of the excerpted East Slavic loanwords and also of the select derivatives, as an intralinguistic abstraction from phenomena such as irrelevant spelling variation are found across dictionaries. A first draft of such a list of 'metavariants' is constructed, separately for the three East Slavic languages, by simply merging all graphemically identical word forms.[2] The editors may freely alter the set of variants and change the mappings of lexicographically attested forms of these variants. For the different variant forms, quotations and dates must be specified as above.
- A set of *symbolic maps* specifying, for each ordered pair (L_1, L_2) of the five languages involved, whether the borrowing history for a certain subset of 'metavariants' or 'metasenses' involved, or possibly involved, a *direct* borrowing from L_1 into L_2 or not.
- A *philological commentary* on the borrowing history of the loanwords covered in the entry, elaborating on the summary representation in the maps.

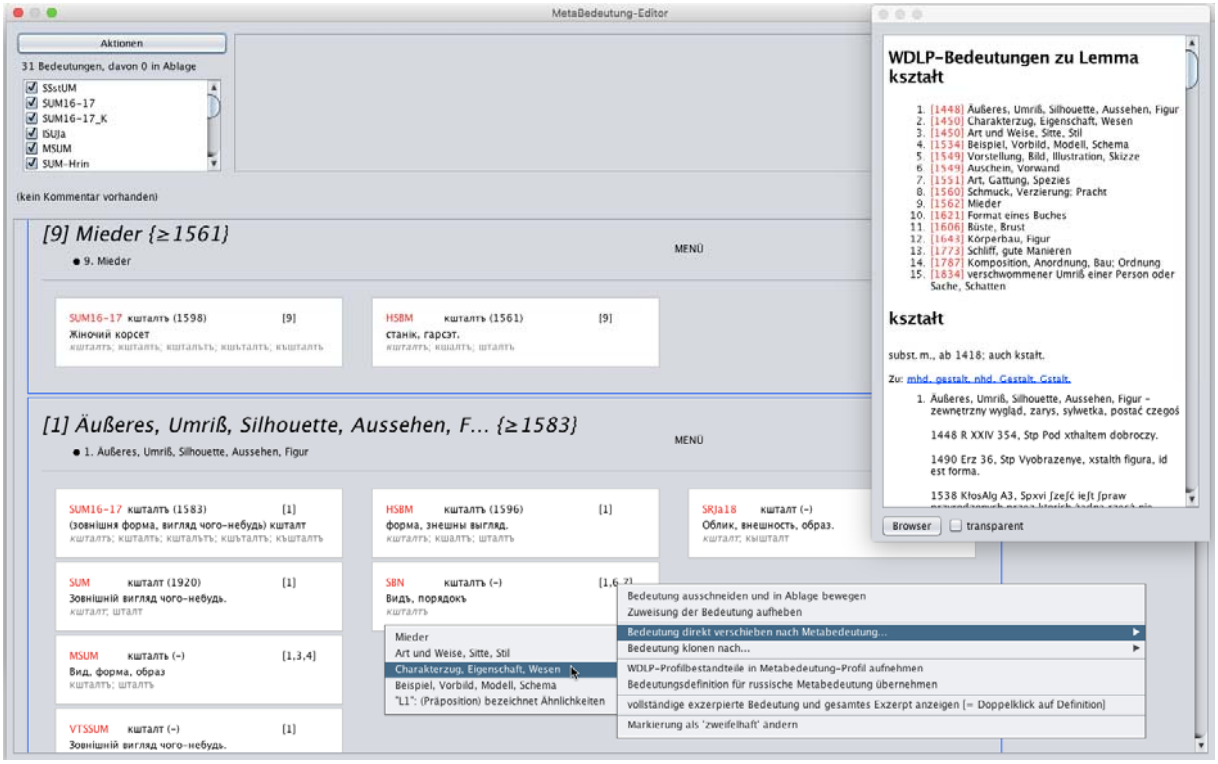

Figure 2: Screenshot of the tool for editing 'metasenses'.

The target entries will be published in the *Lehnwortportal Deutsch* (lwp.ids-mannheim.de), a web portal for loanword dictionaries with German as common donor language; see Meyer 2014c for more details on the portal architecture.

3 Implementation details

The system is based on a cryptographically secured collaborative server/client infrastructure. All programs run cross-platform on any recent standard JVM platform (Java 7 or higher). Many of the client tools, in particular those for excerpting, are part of a standard Swing desktop application written in Java and Groovy; most of the compilation tools run in the browser and are powered by a

[2] Word-final back yer letters ('hard signs') are omitted.

web server running embedded in the desktop application. The application communicates via HTTP with with the web service (implemented with the Java-based Play webframework) in order to modify data stored in a relational database (Oracle 11g). The web service takes care of reporting, backup tasks, and guarantees mutually exclusive access to excerpts and newly compiled entries.
Internally, the tools operate with an underlying object-oriented data model, but all documents are stored on the backend as XML documents using standard (de)serialization (see Meyer 2014b for details); the relational representation used for local SQL queries is generated using a Hibernate-based object-relational mapper. Cross-document data (such as the 'metasenses') is stored in separate relational tables and uses UUIDs for addressing XML fragments.

4 Status of the Project

The wdlpOst system has been in productive project use for excerpting data from the source dictionaries since mid-2014. The toolset is being updated and expanded continually; the 'metasense' tool has been in use since mid-2015, whereas most of the compilation tools will be ready for use in mid-2016. Open-source publication of a sufficiently generalized version of the various tools is beyond the limited resources of the current project, but planned as part of an already submitted follow-up project proposal.

5 References

de Vincenz, A., Hentschel, G. (eds.) (2010). *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts*. Oldenburg: BIS-Verlag. Accessed at: http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp [04/25/2016].

Haspelmath, M., Tadmor, U. (eds.) (2009). *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Accessed at: http://wold.clld.org [04/25/2016].

Meyer, P. (2014a). Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries. In A. Abel, Ch. Vettori & N. Ralli (eds.) The User in Focus. *Proceedings of the XVI EURALEX International Congress: 15-19 July 2014*. Bolzano/Bozen, Bolzano/Bozen: EURAC research, pp. 1135-1144. Accessed at: http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_088_p_1135.pdf [04/25/2016].

Meyer, P. (2014b). Entlehnungsketten in einem Internetportal für Lehnwörterbücher. IT-Infrastruktur und computerlexikographischer Prozess in einem Projekt zu polnisch vermittelten Germanismen im Ostslavischen. In M. Mann (ed.) *Digitale Lexikographie. Ein- und mehrsprachige elektronische Wörterbücher mit Deutsch: aktuelle Entwicklungen und Analysen*. Hildesheim: Olms, pp. 97-132.

Meyer, P. (2014c). Von XML zum DAG: Der lexikographische Prozess bei der Erstellung eines graphenbasierten Wörterbuchportals. In M. J. Domínguez Vázquez, F. Mollica, M. Nied Curcio (eds.): *Zweisprachige Lexikographie zwischen Translation und Didaktik*. Berlin/Boston: de Gruyter (Lexicographica, Series Maior, 147), pp. 303-321.

Meyer, P. (2015). Aligning word senses and more: tools for creating interlinked resources in historical loanword lexicography. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.) Electronic lexicography in the 21st century: linking lexical data in the digital age. *Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton:

Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 198-210. Accessed at: https://elex.link/elex2015/proceedings/eLex_2015_13_Meyer.pdf [04/25/2016].

Acknowledgements

Most of the lexicographical and philological concepts underlying the software presented in this paper have been devised by the project team at the University of Oldenburg. In particular, I would like to thank Gerd Hentschel and Sabine Ute Anders-Marnowsky for many fruitful discussions and continual support.